Data-driven Manifolds for Outdoor Motion Capture

Gerard Pons-Moll[1], Laura Leal-Taixé[1], Juergen Gall[2,3], Bodo Rosenhahn[1]

[1] Leibniz University, Hannover, Germany
[2] BIWI, ETH Zurich, Switzerland
[3] MPI for Intelligent Systems, Germany

Abstract. Human motion capturing (HMC) from multiview image sequences is an extremely difficult problem due to depth and orientation ambiguities and the high dimensionality of the state space. In this paper, we introduce a novel hybrid HMC system that combines video input with sparse inertial sensor input. Employing an annealing particle-based optimization scheme, our idea is to use orientation cues derived from the inertial input to sample particles from the manifold of valid poses. Then, visual cues derived from the video input are used to weight these particles and to iteratively derive the final pose. As our main contribution, we propose an efficient sampling procedure where the particles are derived analytically using inverse kinematics on the orientation cues. Additionally, we introduce a novel sensor noise model to account for uncertainties based on the von Mises-Fisher distribution. Doing so, orientation constraints are naturally fulfilled and the number of needed particles can be kept very small. More generally, our method can be used to sample poses that fulfill arbitrary orientation or positional kinematic constraints. In the experiments, we show that our system can track even highly dynamic motions in an outdoor environment with changing illumination, background clutter, and shadows.

1 Introduction

Recovering 3D human motion from 2D video footage is an active field of research [21, 3, 7, 10, 33, 37]. Although extensive work on human motion capturing (HMC) from multiview image sequences has been pursued for decades, there are only few works, *e.g.* [15], that handle challenging motions in outdoor scenes.

To make tracking feasible in complex scenarios, motion priors are often learned to constrain the search space [18, 29, 30, 32, 37]. On the downside, such priors impose certain assumptions on the motions to be tracked, thus limiting the applicability of the tracker to general human motions. While approaches exist to account for transitions between different types of motion [2, 5, 11], general human motion is highly unpredictable and difficult to be modeled by pre-specified action classes.

Even under the use of strong priors, video HMC is limited by current technology: depth ambiguities, occlusions, changes in illumination, as well as shadows and background clutter are frequent in outdoor scenes and make state-of-the-art algorithms break down. Using many cameras does not resolve the main difficulty in outdoor scenes, namely extracting reliable image features. Strong lighting conditions also rule out the use of depth cameras. Inertial sensors (IMU) do not suffer from such limitations but they are intrusive by nature: at least 17 units must be attached to the body which poses

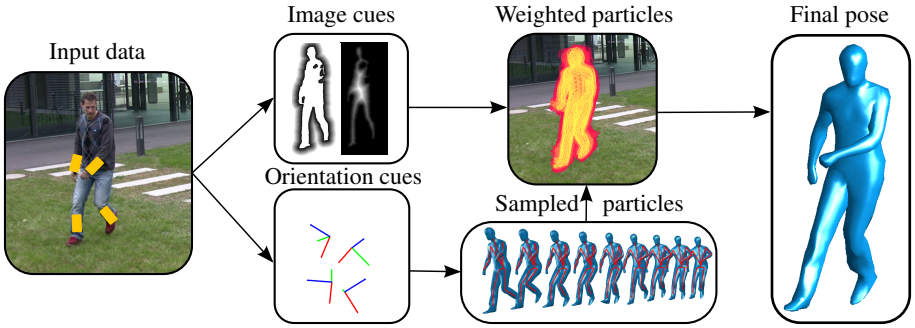


Fig. 1: Orientation cues extracted from inertial sensors are used to efficiently sample valid poses using inverse kinematics. The generated samples are evaluated against image cues in a particle filter framework to yield the final pose.

a problem from bio-mechanical studies and sports sciences. Additionally, IMU's alone fail to measure accurately translational motion and suffer from drift. Therefore, similar to [27, 24, 35], we argue for a hybrid approach where visual cues are supplemented by orientation cues obtained by a small number of additional inertial sensors. While in [35] only arm motions are considered, the focus in [24] is on indoor motions in a studio environment where the cameras and sensors can be very accurately calibrated and the images are nearly noise- and clutter-free. By contrast, we consider full-body tracking in an outdoor setting where difficult lighting conditions, background clutter, and calibration issues pose additional challenges. The work presented here is an extension of our previous article [27]. Here, we extend it and show more results and more implementation details of the proposed approach.

In this paper, we introduce a novel hybrid tracker that combines video input from four consumer cameras with orientation data from five inertial sensors, see Fig. 1. Within a probabilistic optimization framework, we present several contributions that enable robust tracking in challenging outdoor scenarios. Firstly, we show how the high-dimensional space of all poses can be projected to a lower-dimensional manifold that accounts for kinematic constraints induced by the orientation cues. To this end, we introduce an explicit analytic procedure based on Inverse Kinematics (IK). Secondly, by sampling particles from this low-dimensional manifold the constraints imposed by the orientation cues are implictly fulfilled. Therefore, only a small number of particles is needed, leading to a significant improvement in efficiency. Thirdly, we show how to integrate a sensor noise model based on the von Mises-Fisher [8] distribution in the optimization scheme to account for uncertainties in the orientation data. In the experiments, we demonstrate that our approach can track even highly dynamic motions in complex outdoor settings with changing illumination, background clutter, and shadows. We can resolve typical tracking errors such as miss-estimated orientations of limbs and swapped legs that often occur in pure video-based trackers. Moreover, we compare it

with three different alternative methods to integrate orientation data. Finally, we make the challenging dataset and sample code used in this paper available for scientific use[4].

2 Related Work

For solving the high-dimensional pose optimization problem, many approaches rely on local optimization techniques [4, 15, 28], where recovery from false local minima is a major issue. Under challenging conditions, global optimization techniques based on particle filters [7, 10, 38, 26] have proved to be more robust against ambiguities in the data. Thus, we build upon the particle-based annealing optimization scheme described in [10]. Here, one drawback is the computational complexity which constitutes a bottleneck when optimizing in high-dimensional pose spaces.

Several approaches show that constraining particles using external pose information sources can reduce ambiguities [1, 12, 13, 16, 17, 20, 34]. For example, [17] uses the known position of an object a human actor is interacting with and [1, 20] use hand detectors to constrain the pose hypotheses. To integrate such constraints into a particle-based framework, several solutions are possible. Firstly, the cost function that weights the particles can be augmented by additional terms that account for the constraints. Although robustness is added, no benefits in efficiency are achieved, since the dimensionality of the search space is not reduced. Secondly, rejection sampling, as used in [17], discards invalid particles that do not fulfill the constraints. Unfortunately, rejection sampling can be very inefficient and does not scale well with the number of constraints as we will show. Thirdly, approaches such as [9, 12, 19, 34] suggest to explicitly generate valid particles by solving an IK problem on detected body parts. While the proposals in [19, 34] are tailored to deal with depth ambiguities in monocular imagery, [12] relies on local optimization which is not suited for outdoor scenes as we will show. In the context of particle filters, the von Mises-Fisher distribution has been used as prior distribution for extracting white matter fiber pathways from MRI data [40].

In contrast to previous work, our method can be used to sample particles that fulfill arbitrary kinematic constraints by reducing the dimension of the state space. Furthermore, none of the existing approaches perform a probabilistic optimization in a constrained low-dimensional manifold. We introduce an IK based on the *Paden-Kahan* sub-problems and model rotation noise with the von Mises-Fisher distribution.

3 Global Optimization with Sensors

To temporally align and calibrate the input data obtained from a set of uncalibrated and unsynchronized cameras and from a set of orientation sensors, we apply preprocessing steps as explained in Sect. 3.1. Then, we define orientation data within a human motion model (Sect. 3.2) and explain the probabilistic integration of image and orientation cues into a particle-based optimization framework (Sect. 3.3).

[4] http://www.tnt.uni-hannover.de/∼pons/

3.1 Calibration and Synchronization

We recorded several motion sequences of subjects wearing 10 inertial sensors (we used XSens [36]) which we split in two groups of 5: the *tracking sensors* which we use for tracking and the *validation sensors* which we use for evaluation. According to the specifications, the IMU orientation accuracy is around 2° for smooth motions and in abscence of magnetic field. In practice, unfortunatelly, the error is much higher due to different sources of uncertainty, see Sect.4.3. The tracking sensors are placed in the back and the lower limbs and the validation sensors are placed on the chest and the upper limbs. An inertial sensor s measures the orientation of its local coordinate system F_s^S w.r.t. a fixed global frame of reference F^T. All sensors derive the same global frame of reference by merging information from a magnetic field sensor, an accelerometer and a rate gyro. The orientation data is given as a stream of rotation matrices $\mathbf{R}_s^{TS}(t)$ that define the coordinate transform from F_s^S to F^T. In the process of calibrating the camera, the global tracking coordinate system F^T is defined by a calibration cube placed into the recording volume. In order to bring F^I and F^T into correspondence, we carefully place the calibration cube such that the axes of F^T directly correspond to the axes of the known F^I using a compass. Like this, the orientation data $\mathbf{R}_s^{IS}(t)$ also directly maps from the local sensor coordinate system F_s^S to the global tracking coordinate system F^T and we note $\mathbf{R}^{TS} := \mathbf{R}^{IS}$. Note that there might be slight missalignments between the tracking and inertial frame for which we compensate bt introducing a sensor noise model, see Sec. 4.3. In this paper, we refer to the sensor orientations by $\mathbf{R}^{TS}$ and, where appropriate, by using the corresponding quaternion representation $\mathbf{q}^{TS}$. Quaternions generalize complex numbers and can be used to represent 3D rotations the same way as complex numbers can be used to represent planar rotations [31]. The video sequences recorded with four off-the-shelf consumer cameras are synchronized by cross correlating the audio signals as proposed in [15]. Finally, we synchronize the IMU's with the cameras using a clapping motion, which can be detected in the audio data as well as in the acceleration data measured by IMU's.

3.2 Human Motion Model

We model the motion of a human by a skeletal kinematic chain containing $N = 25$ joints that are connected by rigid bones. The global position and orientation of the kinematic chain are parameterized by a twist $\xi_0 \in \mathbb{R}^6$ [22]. A twist is an element of the tangent space of rigid body motions, see [26] for a comprehensive introduction to human body parameterizations. Together with the joint angles $\Theta := (\theta_1 \dots \theta_N)$, the configuration of the kinematic chain is fully defined by a D=6+N-dimensional vector of pose parameters $\mathbf{x} = (\xi_0, \Theta)$. We now describe the relative rigid motion matrix $\mathbf{G}_i$ that expresses the relative transformation introduced by the rotation in the $i-th$ joint. A joint in the chain is modeled by a location $\mathbf{m}_i$ and a rotation axis ω_i. The exponential map of the corresponding twist $\xi_i = (-\omega_i \times \mathbf{m}_i, \omega_i)$ yields $\mathbf{G}_i$ by

$$\mathbf{G}_i = \exp(\theta_i \widehat{\xi_i}). \tag{1}$$

Let $\mathcal{J}_i \subseteq \{1, \ldots, n\}$ be the ordered set of parent joint indices of the $i-th$ bone. The total rigid motion $\mathbf{G}_i^{TB}$ of the bone is given by concatenating the global transformation matrix $\mathbf{G}_0 = \exp(\widehat{\xi}_0)$ and the relative rigid motions matrices $\mathbf{G}_i$ along the chain by

$$\mathbf{G}_i^{TB} = \mathbf{G}_0 \prod_{j \in \mathcal{J}_i} \exp(\theta_j \widehat{\xi}_j). \tag{2}$$

The rotation part of $\mathbf{G}_i^{TB}$ is referred to as *tracking bone orientation* of the $i-th$ bone. In the standard configuration of the kinematic chain, *i.e.* , the zero pose, we choose the local frames of each bone to be coincident with the global frame of reference F^T. Thus, $\mathbf{G}_i^{TB}$ also determines the orientation of the bone relative to F^T. A surface mesh of the actor is attached to the kinematic chain by assigning every vertex of the mesh to one of the bones. Let $\bar{\mathbf{p}}$ be the homogeneous coordinate of a mesh vertex $\mathbf{p}$ in the zero pose associated to the $i-th$ bone. For a configuration $\mathbf{x}$ of the kinematic chain, the vertex is transformed to $\bar{\mathbf{p}}'$ using $\bar{\mathbf{p}}' = \mathbf{G}_i^{TB} \bar{\mathbf{p}}$.

3.3 Optimization Procedure

If several cues are available, *e.g.* image silhouettes and sensor orientation $\mathbf{z} = (\mathbf{z}^{\text{im}}, \mathbf{z}^{\text{sens}})$, the likelihood is commonly factored in two independent terms:

$$\arg\max_{\mathbf{x}} \; p(\mathbf{x}|\mathbf{z}^{\text{im}}, \mathbf{z}^{\text{sens}}) = p(\mathbf{z}^{\text{im}}|\mathbf{x}) p(\mathbf{z}^{\text{sens}}|\mathbf{x}) p(\mathbf{x}) \tag{3}$$

where it is assumed that the measurements $\mathbf{z}^{\text{im}}$ and $\mathbf{z}^{\text{sens}}$ are conditionaly independent given that the pose $\mathbf{x}$ is known. The human pose $\mathbf{x}$ can then be found by minimizing the negative $\log$-likelihood which yields a weighted combination of cost functions for both terms as in [24]. Since in outdoor scenarios the sensors are not perfectly calibrated and the observations are noisy, fine tuning of the weighting parameters would be necessary to achieve good performance. Furthermore, the orientation information is not used to reduce the state space, and thus the optimization cost and ambiguities. Hence, we propose a different probabilistic formulation of the problem:

$$p(\mathbf{x}|\mathbf{z}^{\text{im}}, \mathbf{z}^{\text{sens}}) = \frac{p(\mathbf{z}^{\text{im}}, \mathbf{z}^{\text{sens}}|\mathbf{x}) p(\mathbf{x})}{p(\mathbf{z}^{\text{im}}, \mathbf{z}^{\text{sens}})} = \frac{p(\mathbf{z}^{\text{im}}|\mathbf{x}) p(\mathbf{z}^{\text{sens}}|\mathbf{x}) p(\mathbf{x})}{p(\mathbf{z}^{\text{im}}) p(\mathbf{z}^{\text{sens}})} \tag{4}$$

where we assumed independence between sensors and using

$$p(\mathbf{x}|\mathbf{z}^{\text{sens}}) = \frac{p(\mathbf{z}^{\text{sens}}|\mathbf{x}) p(\mathbf{x})}{p(\mathbf{z}^{\text{sens}})}$$

we obtain the following factorized posterior

$$p(\mathbf{x}|\mathbf{z}^{\text{im}}, \mathbf{z}^{\text{sens}}) \propto p(\mathbf{z}^{\text{im}}|\mathbf{x}) p(\mathbf{x}|\mathbf{z}^{\text{sens}}). \tag{5}$$

that can be optimized globally and efficiently. We disregard the normalization factor $p(\mathbf{z}^{\text{im}})$ since it does not depend on the pose $\mathbf{x}$. The weighting function $p(\mathbf{z}^{\text{im}}|\mathbf{x})$ can be modeled by any image-based likelihood function. Our proposed model of $p(\mathbf{x}|\mathbf{z}^{\text{sens}})$,

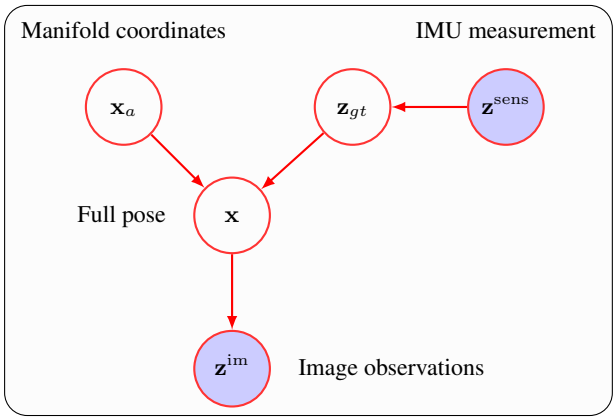


Fig. 2: Graphical model of the approach. The measurements $\mathbf{z}^{\text{im}}$ and $\mathbf{z}^{\text{sens}}$ are shown as shaded nodes because they are observable during inference. The manifold coordinates, $\mathbf{x}_a$, the full state pose $\mathbf{x}$ and the true orientations $\mathbf{z}_{gt}$ are hidden. To infer the full state pose $\mathbf{x}$ we optimize the manifold coordinates and marginalize out $\mathbf{z}_{gt}$. To integrate out $\mathbf{z}_{gt}$, we assume it follows a von-Mises-Fisher distribution with mean direction $\boldsymbol{\mu} = \mathbf{z}^{\text{sens}}$.

as introduced in Sect. 4, integrates uncertainties in the sensor data and constrains the poses to be evaluated to a lower dimensional manifold. For single frame pose estimation, optimization is typically performed by importance sampling, *i.e.* sampling from the prior $p(\mathbf{x})$ and weighting by the likelihood function $p(\mathbf{z}^{\text{im}}|\mathbf{x})$. The problem with this is that the prior is broad compared to $p(\mathbf{z}^{\text{im}}|\mathbf{x})$ that is peaky and typically multi-valued. By drawing proposals directly from $p(\mathbf{x}|\mathbf{z}^{\text{sens}})$ we are effectively reducing the number of wasted samples, *i.e.* we are concetrating samples on the likelihood region. For optimization, we use the method proposed in [10]; the implementation details are given in Sect. 4.4.

4 Manifold Sampling

Assuming that the orientation data $\mathbf{z}^{\text{sens}}$ of the N_s orientation sensors is accurate and that each sensor has 3 DoF that are not redundant [5], the D dimensional pose $\mathbf{x}$ can be reconstructed from a lower dimensional vector $\mathbf{x}_a \in \mathbb{R}^d$ where $d = D - 3N_s$. In our experiments, a 31 DoF model can be represented by a 16 dimensional manifold using 5 inertial sensors as shown in Fig. 5 (a). The mapping is denoted by $\mathbf{x} = g^{-1}(\mathbf{x}_a, \mathbf{z}^{\text{sens}})$ and is described in Sect. 4.1. In this setting, Eq. (3) can be rewritten as

$$\arg\max_{\mathbf{x}_a} \; p\left(\mathbf{z}^{\text{im}}|g^{-1}(\mathbf{x}_a, \mathbf{z}^{\text{sens}})\right). \tag{6}$$

[5] Since the sensors are placed in different body parts they are not redundant because they explain different DoF in the kinematic chain

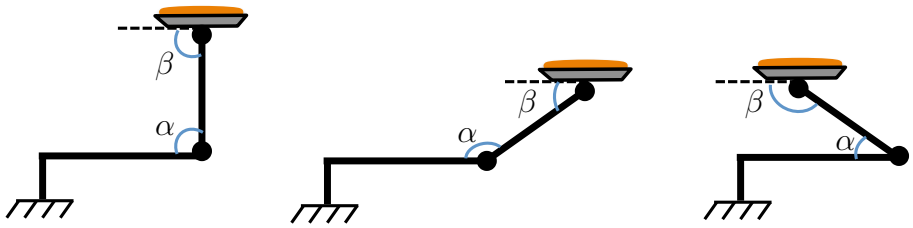


Fig. 3: Toy example to illustrate our idea to sample from lower dimensional manifolds. For this simple kinematic chain the state vector has 2 DoF, $\mathbf{x} = (\alpha, \beta)$. If we impose the constraint that the cake plate must be perpendicular to the ground the true state vector has dimensionality 1. The constraint is $\alpha + \beta = \pi$ and therefore the state vector can be re-parameterized as $\mathbf{x} = (\alpha, \pi - \alpha)$. For the problem of human pose estimation however the constraints are non-linear and therefore re-parametrization is achieved by solving small Inverse Kinematics subproblems.

Since the orientation data $\mathbf{z}^{\text{sens}}$ is not always accurate due to sensor noise and calibration errors, we introduce a term $p(\mathbf{z}_{gt}^{\text{sens}}|\mathbf{z}^{\text{sens}})$ that models the sensor uncertainty, *i.e.* , the probability of the true orientation $\mathbf{z}_{gt}^{\text{sens}}$ given the sensor data $\mathbf{z}^{\text{sens}}$. We assume the conditional probability $p(\mathbf{z}_{gt}^{\text{sens}}|\mathbf{z}^{\text{sens}})$ follows a *von-Mises Fisher* distribution and it is described in detail Sect. 4.3. Hence, we get the final objective function:

$$\arg\max_{\mathbf{x}_a} \int p\left(\mathbf{z}^{\text{im}}|g^{-1}(\mathbf{x}_a, \mathbf{z}_{gt}^{\text{sens}})\right) p\left(\mathbf{z}_{gt}^{\text{sens}}|\mathbf{z}^{\text{sens}}\right) d\mathbf{z}_{gt}^{\text{sens}}. \tag{7}$$

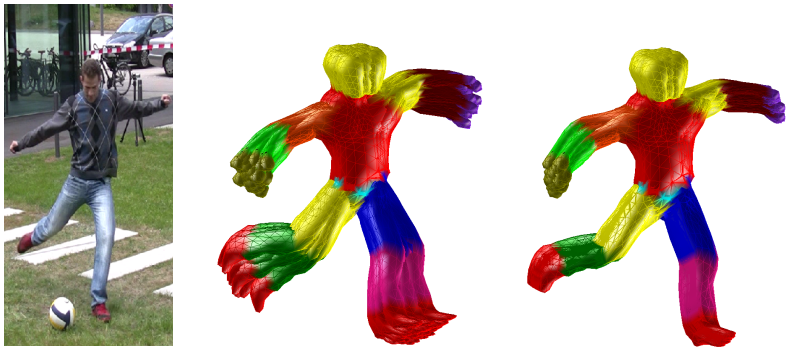
where we marginalize out the sensor noise and optimize the manifold coordinates. The integral can be approximated by importance sampling, *i.e.* , drawing particles from $p(\mathbf{z}_{gt}^{\text{sens}}|\mathbf{z}^{\text{sens}})$ and weighting them by $p(\mathbf{z}^{\text{im}}|\mathbf{x})$. Consequently, we can efficiently concentrate the search space in the neighborhood region of a low dimensional manifold. In addition, we can guarantee that the kinematic constraints are satisfied.

4.1 Inverse Kinematics using Inertial Sensors

For solving Eq. (7), we derive an analytical solution for the map $g : \mathbb{R}^D \mapsto \mathbb{R}^{D-3N_s}$ and its inverse g^{-1}. Here, g projects $\mathbf{x} \in \mathbb{R}^D$ to a lower dimensional space and its inverse function g^{-1} uses the sensor orientations and the coordinates in the lower dimensional space $\mathbf{x}_a \in \mathbb{R}^{D-3N_s}$ to reconstruct the parameters of the full pose, *i.e.* ,

$$g(\mathbf{x}) = \mathbf{x}_a \qquad g^{-1}(\mathbf{x}_a, \mathbf{z}^{\text{sens}}) = \mathbf{x}. \tag{8}$$

To derive a set of minimal coordinates, we observe that given the full set of parameters $\mathbf{x}$ and the kinematic constraints placed by the sensor orientations, a subset of these parameters can be written as a function $f(\cdot)$ of the others, see Fig. 3 for an intuitive illustration. Specifically, the full set of parameters is decomposed into a set of *active parameters* $\mathbf{x}_a$ which we want to optimize according to Eq. (7) and a set of *passive*



(a) (b) (c)

Fig. 4: Manifold Sampling: **(a)** Original image. **(b)** Full space sampling. **(c)** Manifold sampling. Note that the generated samples in **(c)** have parallel end-effector orientations because they satisfy the constraints and uncertainty is therefore reduced.

parameters $\mathbf{x}_p$ that can be derived from the constraint equations and the active set. Writing the state as $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_p)$ with $\mathbf{x}_a \in \mathbb{R}^d$ and $\mathbf{x}_p \in \mathbb{R}^{D-d}$ we have

$$f(\mathbf{x}_a, \mathbf{z}^{\text{sens}}) = \mathbf{x}_p \qquad \Longrightarrow \qquad g^{-1}(\mathbf{x}_a, \mathbf{z}^{\text{sens}}) = (\mathbf{x}_a, f(\mathbf{x}_a, \mathbf{z}^{\text{sens}})). \tag{9}$$

Thereby, the direct mapping g is trivial since from the full set only the active parameters are retained. The inverse mapping g^{-1} can be found by solving *inverse kinematics* (IK) sub-problems. Several choices for the decomposition into active and passive set are possible. To guarantee the existence of solution for all cases, we choose the passive parameters to be the set of 3 DoF joints that lie on the kinematic branches where a sensor is placed. In our experiments using 5 sensors, we choose the passive parameters to be the two shoulder joints, the two hips and the root joint adding up to a total of 15 parameters which corresponds to $3N_s$ constraint equations, see Fig. 5 (a). Hence, the passive parameters consist of N_s triplets of joint angles $\mathbf{x}_p = (\theta_{j_1}, \theta_{j_2}, \theta_{j_3})^T$, $j \in \{1 \dots N_s\}$ with corresponding rotation matrices $\mathbf{R}_j$. Since each sensor $s \in \{1 \dots N_s\}$ is rigidly attached to a bone, there exists a constant rotational offset $\mathbf{R}_s^{SB}$ between the i-th bone and the local coordinate system F_s^S of the sensor attached to it. This offset can be computed from the tracking bone orientation $\mathbf{R}_{i,0}^{TB}$ in the first frame and the sensor orientation $\mathbf{R}_{s,0}^{TS}$

$$\mathbf{R}_s^{SB} = (\mathbf{R}_{s,0}^{TS})^T \mathbf{R}_{i,0}^{TB}. \tag{10}$$

At each frame t, we obtain *sensor bone orientations* $\mathbf{R}_{s,t}^{TS}\mathbf{R}_s^{SB}$ by applying the rotational offset. In the absence of sensor noise, it is desired to enforce that the tracking bone orientation and the sensor bone orientation are equal:

$$\mathbf{R}_{i,t}^{TB} = \mathbf{R}_{s,t}^{TS}\mathbf{R}_s^{SB} \tag{11}$$

In Sect. 4.3 we show how to deal with noise in the measurements. Let $\mathbf{R}_j$ be the relative rotation of the j-th joint given by the rotational part of Eq. (1). The relative rotation $\mathbf{R}_j$

associated with the passive parameters can be isolated from Eq. (11). To this end, we expand the tracking bone orientation $\mathbf{R}_{i,t}^{TB}$ to the product of 3 relative rotations[6] $\mathbf{R}_j^p$, the total rotation motion of parent joints in the chain, $\mathbf{R}_j$, the unknown rotation of the joint associated with the passive parameters, and $\mathbf{R}_j^c$, the relative motion between the j-th joint and the i-th joint where the sensor is placed:

$$\mathbf{R}_j^p \mathbf{R}_j \mathbf{R}_j^c = \mathbf{R}_s^{TS} \mathbf{R}_s^{SB} \tag{12}$$

Note that $\mathbf{R}_j^p$ and $\mathbf{R}_j^c$ are constructed from the active set of parameters $\mathbf{x}_a$ using the product of exponentials formula (2). From Eq. (12), we obtain the relative rotation matrix

$$\mathbf{R}_j = (\mathbf{R}_j^p)^T \mathbf{R}_s^{TS} \mathbf{R}_s^{SB} (\mathbf{R}_j^c)^T. \tag{13}$$

Having $\mathbf{R}_j$ and the known fixed rotation axes $\omega_{j_1}, \omega_{j_2}, \omega_{j_3}$ of the j-th joint, the rotation angles $\theta_{j_1}, \theta_{j_2}, \theta_{j_3}$, *i.e.* , the passive parameters, must be determined such that

$$\exp(\theta_{j_1} \widehat{\omega}_{j_1}) \exp(\theta_{j_2} \widehat{\omega}_{j_2}) \exp(\theta_{j_3} \widehat{\omega}_{j_3}) = \mathbf{R}_j. \tag{14}$$

This problem can be solved by decomposing it into sub-problems [23], see Sec. 4.2. By solving these sub-problems for every sensor, we are able to reconstruct the full state $\mathbf{x}$ using only a subset of the parameters $\mathbf{x}_a$ and the sensor measurements $\mathbf{z}^{\text{sens}}$. In this way, the inverse mapping $g^{-1}(\mathbf{x}_a, \mathbf{z}_{\text{sens}}) = \mathbf{x}$ is fully defined and we can efficiently sample from the manifold, see Fig. 4.

4.2 Paden-Kahan Subproblems

We are interested in solving the following problem:

$$\exp(\theta_1 \widehat{\omega}_1) \exp(\theta_2 \widehat{\omega}_2) \exp(\theta_3 \widehat{\omega}_3) = \mathbf{R}_j. \tag{15}$$

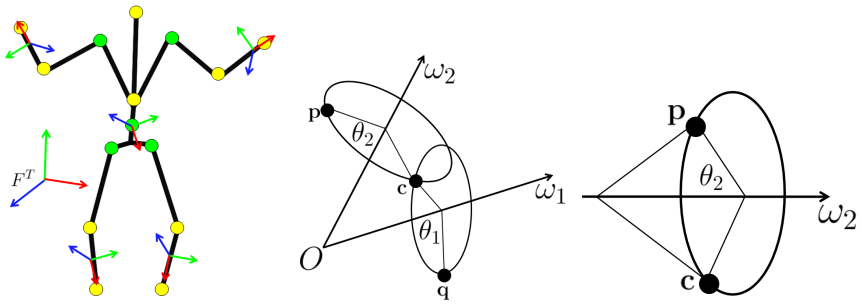
This problem can be solved by decomposing it into sub-problems as proposed in [23]. A comprehensive description of the Paden-Kahan subproblems applied to several inverse kinematic problems can also be found in [22]. The basic technique for simplification is to apply the kinematic equations to specific points. By using the property that the rotation of a point on the rotation axis is the point itself, we can pick a point $\mathbf{p}$ on the third axis ω_3 and apply it to both sides of Eq. (15) to obtain

$$\exp(\theta_1 \widehat{\omega}_1) \exp(\theta_2 \widehat{\omega}_2) \mathbf{p} = \mathbf{R}_j \mathbf{p} = \mathbf{q} \tag{16}$$

which is known as the *Paden-Kahan sub-problem 2*. For our problem the 3 rotation axes intersect at the same joint location. Consequently, since we are only interested in the orientations, we can translate the joint location to the origin $\mathbf{q}_j = O = (0,0,0)^T$. In this way, any point $\mathbf{p} = \lambda \omega_3$ with $\lambda \in \mathbb{R}$, $\lambda \neq 0$ is a valid choice for $\mathbf{p}$. Eq. (16) can decomposed in two subproblems

$$\exp(\theta_2 \widehat{\omega}_2) \mathbf{p} = \mathbf{c} \quad \text{and} \quad \exp(-\theta_1 \widehat{\omega}_1) \mathbf{q} = \mathbf{c}, \tag{17}$$

[6] The temporal index t is omitted for the sake of clarity



(a) (b) (c)

Fig. 5: Inverse Kinematics: **(a)** decomposition into active (yellow) and passive (green) parameters. Paden-Kahan sub-problem 2 **(b)** and sub-problem 1 **(c)**.

where $\mathbf{c}$ is the intersection point between the circles created by the rotating point $\mathbf{p}$ around axis ω_2 and the point $\mathbf{q}$ rotating around axis ω_1 as shown in Fig. 5 (b). In order for Eq. (17) to have a solution, the points $\mathbf{p}$, $\mathbf{c}$ must lie in the same plane perpendicular to ω_2, and $\mathbf{q}$, $\mathbf{c}$ must lie in the same plane perpendicular to ω_1. This implies that the projection of the position vectors [7] $\mathbf{p}, \mathbf{c}, \mathbf{q}$ onto the span of ω_1, ω_2 respectively must be equal, see Fig. 6

$$\omega_2^T \mathbf{p} = \omega_2^T \mathbf{c} \quad \text{and} \quad \omega_1^T \mathbf{q} = \omega_1^T \mathbf{c} \tag{18}$$

Additionally, the norm of a vector is preserved after rotation and therefore

$$\|\mathbf{p}\| = \|\mathbf{c}\| = \|\mathbf{q}\| \tag{19}$$

Since ω_1 and ω_2 are not parallel, the vectors $\omega_1, \omega_2, \omega_1 \times \omega_2$ form a basis that span $\mathbb{R}^3$. Hence, we can write $\mathbf{c}$ in the new basis as

$$\mathbf{c} = \alpha\omega_1 + \beta\omega_2 + \gamma(\omega_1 \times \omega_2) \tag{20}$$

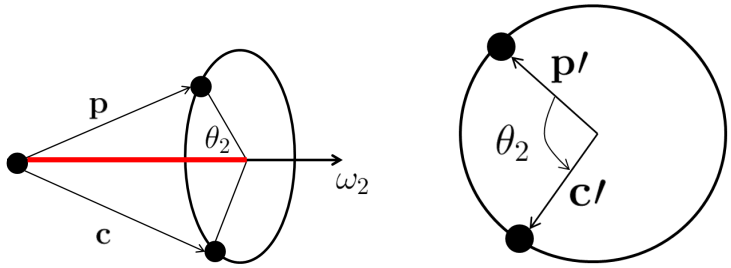
where α, β, γ are the new coordinates of $\mathbf{c}$. Now, using the fact that $\omega_2 \perp \omega_1 \times \omega_2$ and $\omega_1 \perp \omega_1 \times \omega_2$, we can substitute Eq. (20) into Eq. (18) to obtain a system of two equations with two unknowns (α, β)

$$\begin{aligned} \omega_2^T \mathbf{p} &= \alpha \omega_2^T \omega_1 + \beta \\ \omega_1^T \mathbf{q} &= \alpha + \beta \omega_1^T \omega_2 \end{aligned} \tag{21}$$

from which we can isolate the first two coordinates of $\mathbf{c}$

$$\begin{aligned} \alpha &= \frac{(\omega_1^T \omega_2)\omega_2^T \mathbf{p} - \omega_1^T \mathbf{q}}{(\omega_1^T \omega_2)^2 - 1} \\ \beta &= \frac{(\omega_1^T \omega_2)\omega_1^T \mathbf{q} - \omega_2^T \mathbf{p}}{(\omega_1^T \omega_2)^2 - 1} \end{aligned} \tag{22}$$

[7] Since we translated the joint location to the origin we can consider the points as vectors with origin at the joint location $\mathbf{q}_j$.



(a) (b)

Fig. 6: Paden-Kahan subproblem 1: **(a)** the projection length of $\mathbf{p}$ and $\mathbf{c}$ onto ω_2 must be equal, **(b)** the projection of the vectors $\mathbf{p}$ and $\mathbf{c}$ onto the orthogonal plane to the rotation axes ω_2

From Eq. (19) and Eq. (20) we can write

$$\|\mathbf{p}\|^2 = \|\mathbf{c}\|^2 = \alpha^2 + \beta^2 + 2\alpha\beta\omega_1^T\omega_2 + \gamma^2\|\omega_1 \times \omega_2\|^2 \tag{23}$$

and obtain the third coordinate γ as

$$\gamma^2 = \frac{\|\mathbf{p}\|^2 - \alpha^2 - \beta^2 - 2\alpha\beta\omega_1^T\omega_2}{\|\omega_1 \times \omega_2\|^2} \tag{24}$$

This last equation has no solution when the circles do not intersect, one solution when the circles are tangential and two solutions when the circles intersect at two points. For our choice of decomposition, the passive parameters correspond to $3DoF$ joints which are modeled as 3 concatenated revolute joints whose axis are mutually orthogonal. Therefore, there always exists a solution [22]. We note that the inverse kinematic solutions presented here are also valid for other decompositions, *e.g.* one could choose as passive parameters two rotation axes of the shoulder joint and one rotation axis of the elbow joints. However, the existence of solution should then be checked during the sampling process. Once we have the new coordinates (α, β, γ) we can obtain the intersection point $\mathbf{c}$ in the original coordinates using equation Eq. (20). Thereafter, Eq. (17) can be decomposed into two problems of the form

$$\begin{aligned} \exp(\theta_2\widehat{\omega}_2)\mathbf{p} &= \mathbf{c} \\ \exp(-\theta_1\widehat{\omega}_1)\mathbf{q} &= \mathbf{c} \end{aligned} \tag{25}$$

which simplifies to finding the rotation angle about a fixed axis that brings a point $\mathbf{p}$ to a second one $\mathbf{c}$, which is known as *Paden-Kahan sub-problem 1*

$$\exp(\theta_2\widehat{\omega}_2)\mathbf{p} = \mathbf{c}. \tag{26}$$

This problem has a solution when the projections of the vectors $\mathbf{p}$ and $\mathbf{c}$ onto the orthogonal plane to ω_2 have equal lengths. Let $\mathbf{p}'$ and $\mathbf{c}'$ be the projections of $\mathbf{p}, \mathbf{c}$ onto the plane perpendicular to ω_2, see Fig. 6,

$$\mathbf{p}' = \mathbf{p} - \omega_2\omega_2^T\mathbf{p} \quad \text{and} \quad \mathbf{c}' = \mathbf{c} - \omega_2\omega_2^T\mathbf{c}. \tag{27}$$

If the projections have equal lengths $\|\mathbf{p}'\| = \|\mathbf{c}'\|$ then the problem is as simple as finding the angle between the two vectors

$$\begin{aligned} \omega_2^T(\mathbf{p}' \times \mathbf{c}') &= \sin\theta_2 \|\mathbf{p}'\|\|\mathbf{c}'\| \\ \mathbf{p}' \cdot \mathbf{c}' &= \cos\theta_2 \|\mathbf{p}'\|\|\mathbf{c}'\| \end{aligned} \tag{28}$$

By dividing the equations we finally obtain the rotation angle using the arc tangent

$$\theta_2 = \text{atan2}(\omega_2^T(\mathbf{p}' \times \mathbf{c}'), \mathbf{p}' \cdot \mathbf{c}'). \tag{29}$$

We can find θ_1 using the same procedure. Finally, θ_3 is obtained from Eq. (15) after substituting θ_1 and θ_2

$$\exp(\theta_3\widehat{\omega}_3) = \exp(\theta_1\widehat{\omega}_1)^T \exp(\theta_2\widehat{\omega}_2)^T \mathbf{R}_j = \mathbf{R} \tag{30}$$

where the rotation matrix $\mathbf{R}$ is known. The rotation angle θ_3 satisfies

$$2\cos\theta_3 = (\text{trace}(\mathbf{R}) - 1) \tag{31}$$

$$2\sin\theta_3 = \omega_3^T \mathbf{r} \tag{32}$$

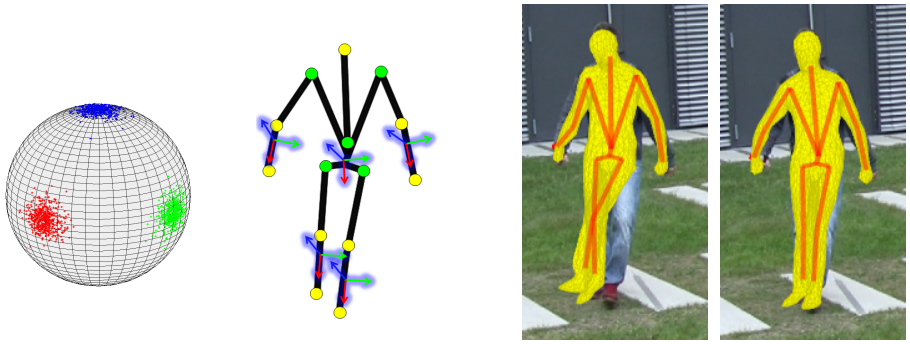
where $\mathbf{r} = (\mathbf{R}_{32} - \mathbf{R}_{23}, \mathbf{R}_{13} - \mathbf{R}_{31}, \mathbf{R}_{21} - \mathbf{R}_{12})$ (page 584 of [14]). Finally, the rotation angle θ_3 can be computed from $\cos\theta_3$ and $\sin\theta_3$ using atan2. By solving these sub-problems for every sensor, we are able to reconstruct the full state $\mathbf{x}$ using only a subset of the parameters $\mathbf{x}_a$ and the sensor measurements $\mathbf{z}^{\text{sens}}$. The good property of this geometric algorithms for solving inverse kinematics is that they are numerically very stable. More importantly, the same principle can be applied to solve more complex IK problems involving a number of positional and orientational constraints.

4.3 Sensor Noise Model

In practice, perfect alignment and synchronization of inertial and video data is not possible. In fact, there are at least four sources of uncertainty in the inertial sensor measurements, namely inherent sensor noise from the device, temporal unsynchronization with the images, small alignment errors between the tracking coordinate frame F^T and the inertial frame F^I, and errors in the estimation of $\mathbf{R}_s^{SB}$. Hence, we introduce a noise model $p(\mathbf{z}_{gt}^{\text{sens}}|\mathbf{z}^{\text{sens}})$ in our objective function (7). Rotation errors are typically modeled by assuming that the measured rotations are distributed according to a Gaussian in the tangent spaces which is implemented by adding Gaussian noise v^i on the parameter components, *i.e.* , $\tilde{\mathbf{x}}_j = \mathbf{x}_j + v^i$. The topological structure of the elements, a 3-sphere S^3 in case of quaternions, is therefore ignored. The *von Mises-Fisher* (MF) distribution models errors of elements that lie on a unit sphere S^{p-1} [8] and is defined as

$$f_p(\mathbf{x}; \boldsymbol{\mu}, \kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{d/2-1}(\kappa)} \exp(\kappa \boldsymbol{\mu}^T \mathbf{x}) \tag{33}$$

where I_v denotes the modified Bessel function of the first kind, $\boldsymbol{\mu}$ is the mean direction, and κ is a concentration parameter that determines the dispersion form the true position.



(a) (b) (c) (d)

Fig. 7: Sensor noise model. **(a)** Points disturbed with rotations sampled from a von Mises-Fisher distribution. **(b)** The orientation of the particles can deviate from the sensor measurements. Tracking without **(c)** and with **(d)** sensor noise model.

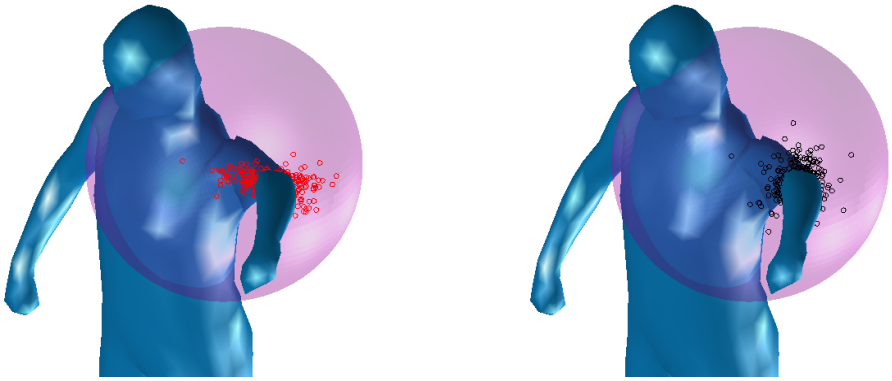
The distribution is illustrated in Fig. 7. For our problem, $p = 4$ and thus the elements $\mathbf{x}$ are quaternions. Therefore, on the one hand samples of the MF disutribution are quaternions whose corresponding axis of rotation are uniformly distributed in all directions. On the other hand, the sample concentration decays with the angle of rotation. To see this, observe that the distribution can be expressed as a function of the angular rotation θ from the mean $\boldsymbol{\mu}$ where we replaced the inner product $\boldsymbol{\mu}^T\mathbf{x}$ by $\cos\left(\frac{\theta}{2}\right)$ (the inner prodcut between two quaternions results in $\cos(\frac{\theta}{2})$, where θ is the geodesic angle distance between rotations).

In order to approximate the integral in Eq. (7) by importance sampling, we use the method proposed in [39] to draw samples $\mathbf{q}_w$ from the von Mises-Fisher distribution with $p = 4$ and $\boldsymbol{\mu} = (1, 0, 0, 0)^T$, which is the quaternion representation of the identity. We use a fixed dispersion parameter of $\kappa = 1000$. The sensor quaternions are then rotated by the random samples $\mathbf{q}_w$:

$$\tilde{\mathbf{q}}_s^{TS} = \mathbf{q}_s^{TS} \circ \mathbf{q}_w \tag{34}$$

where $\circ$ denotes quaternion multiplication. In this way, for every particle, samples $\tilde{\mathbf{q}}_s^{TS}$ are drawn from $p(\mathbf{z}_{gt}^{\text{sens}}|\mathbf{z}^{\text{sens}})$ using Eq. (34) obtaining a set of distributed measurements $\tilde{\mathbf{z}}^{\text{sens}} = \left(\tilde{\mathbf{q}}_1^{TS} \dots \tilde{\mathbf{q}}_{N_s}^{TS}\right)$. This can be interpreted as the analogous of additive Gaussian Noise where $\mathbf{q}_w$ is a rotation noise sample. Thereafter, the full pose is reconstructed from the newly computed orientations with $g^{-1}(\mathbf{x}_a, \tilde{\mathbf{z}}^{\text{sens}})$ as explained in Sect. 4.1 and weighted by $p(\mathbf{z}^{\text{im}}|\mathbf{x})$.

In Fig. 8, we compare the inverse kinematic solutions of 500 samples $i \in \{1 \dots 500\}$ by simply adding Gaussian noise *only* on the passive parameters $\{g^{-1}(\mathbf{x}_a, \mathbf{z}^{\text{sens}}) + \mathbf{v}^i\}_i$ and by modeling sensor noise with the von Mises-Fisher distribution $\{g^{-1}(\mathbf{x}_a, \tilde{\mathbf{z}}^{sens,i})\}_i$. For the generated samples, we fixed the vector of manifold coordinates $\mathbf{x}_a$ and we used equivalent dispersion parameters for both methods. To visualize the reconstructed poses we only show, for each sample, the elbow location represented as a point in the sphere. This example shows that simply adding Gaussian noise on the parameters is biased



(a) (b)

Fig. 8: Sensor noise model. 500 samples of the IK elbow location are shown as points using: **(a)** added Gaussian noise and **(b)** noise from the von Mises-Fisher distribution.

towards one direction that depends on the current pose $\mathbf{x}$. By contrast, the samples using von Mises-Fisher are uniformly distributed in all directions and the concentration decays with the angular error from the mean. Note, however, that Fig. 8 is a 3D visualization, in reality the bone orientations of the reconstructed poses should be visualized as points in a 3-sphere S^3.

$$f_p(\theta;\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{d/2-1}(\kappa)} \exp\left(\kappa \cos\left(\frac{\theta}{2}\right)\right) \tag{35}$$

4.4 Implementation Details

To optimize Eq. (7), we have implemented ISA (Interacted Simulated Annealing), the global optimization approach that has been proposed in [10] and use only the first stage of the algorithm, *i.e.* we do not locally optimize. ISA is based on simulated annealing which is a stochastic optimization technique to locate a good approaximation of the global optimum of a cost function in a large search space. In the remainder of this paper we will use the term global optimization whenever ISA was used for optimization to make the distinction with local optimization methods. As cost function, we use the silhouette and color terms

$$V(\mathbf{x}) = \lambda_1 V_{silh}(\mathbf{x}) + \lambda_2 V_{app}(\mathbf{x}) \tag{36}$$

with the setting $\lambda_1 = 2$ and $\lambda_2 = 40$. Although a good likelihood model is essential for good performance, it is not the focus of our work and we refer the interested reader to [26] for more details. During tracking, the initial particles $\{\mathbf{x}_a^i\}_i$ are predicted from the particles in the previous frame using a 3rd order autoregression and projected to the low-dimensional manifold using the mapping g; see Sect. 4.1. The optimization is

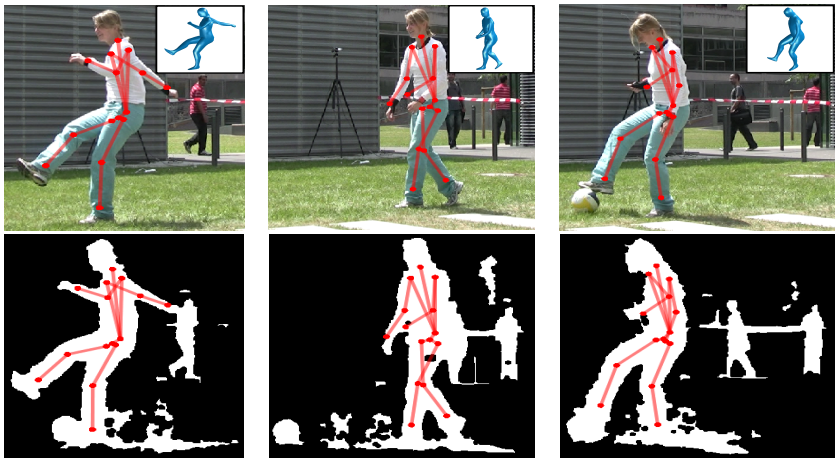


Fig. 9: Tracking with background clutter.

performed only over the active parameters $\mathbf{x}_a \in \mathbb{R}^{D-3N_s}$, *i.e.* , the diffusion step is performed in $\mathbb{R}^{D-3N_s}$. Specifically, diffusion is performed with a Gaussian kernel with zero mean and covariance matrix

$$\Sigma_{a,k} = \frac{\alpha_\Sigma}{N-1}\left(\rho\mathbf{I} + \sum_i^N (\mathbf{x}_{a,k}^{(i)} - \mu_{a,k})(\mathbf{x}_{a,k}^{(i)} - \mu_{a,k})^T\right) \tag{37}$$

proportional to the sampling covariance matrix scaled by α_Σ where μ_k is the particle set mean at the current iteration k.

For the weighting step, we use the approach described in Sect. 4.3 to generate a sample $\tilde{\mathbf{z}}^{\text{sens},i}$ from $p(\mathbf{z}_{gt}^{\text{sens}}|\mathbf{z}^{\text{sens}})$ for each particle $\mathbf{x}_a^i$. Consequently, we can map each particle back to the full space using $\mathbf{x}^i = g^{-1}(\mathbf{x}_a^i, \tilde{\mathbf{z}}^{\text{sens,i}})$ and weight it by

$$\pi_k^{(i)} = \exp\left(-\beta_k \cdot V\left(g^{-1}(\mathbf{x}_{a,k}^i, \tilde{\mathbf{z}}^{\text{sens}})\right)\right), \tag{38}$$

where β_k is the inverse temperature of the annealing scheme at iteration k and $V(\cdot)$ is the image cost function defined in Eq. (36). From the obtained set of weighted particles $\{\pi_k^{(i)}, \mathbf{x}_{a,k}^{(i)}\}_{i=1}^N$ we draw a new set of particles with resampling and probability equal to the normalized weights. The weighting, resampling and diffusion step are iterated M times before going to the next frame. In our experiments, we used 15 iterations for optimization. Finally, the pose estimate is obtained from the remaining particle set at the last iteration as

$$\hat{\mathbf{x}}_t = \sum_i \pi_k^{(i)}\, g^{-1}(\mathbf{x}_{a,k}^{(i)}, \tilde{\mathbf{z}}^{sens,i}). \tag{39}$$

The steps of our proposed sampling scheme are outlined in Algorithm 1.

Dynamics: To model the dynamics we use a 3rd order auto-regression using Gaussian Process regression that provides a prediction $\mathbf{x}^{pred}$ and a covariance matrix Σ^{pred} related with the confidence of the prediction. Thereby, the particles from the previous frame are drifted towards the predicted mean $\mathbf{x}^{pred}$ and diffused with a Gaussian kernel with zero mean and covariance Σ^{pred}. In order to obtain the low dimensional particle set, every particle is projected $g(\mathbf{x}_t^i) = \mathbf{x}_{a,t}^{(i)}$ [8]. We note that we do not learn a mapping directly in the low dimensional space since the previous estimates of passive parameters $\mathbf{x}_{p,t-4:t-1}$ are in general also correlated with the active parameters $\mathbf{x}_{a,t}$. The particle set is used as the initial proposal distribution for the first iteration of ISA.

Algorithm 1 Proposed algorithm

Require: number of layers M, number of samples N, initial distribution $\mathcal{L}_0$, sensor orientations $\mathbf{z}^{sens}$, image cost function $V(\cdot)$

Initialize: Draw N initial samples from $\mathcal{L}_0 \rightarrow \boldsymbol{x}_{a,k}^{(i)}$

for layer $k = 0$ to M **do**

1. *MANIFOLD SAMPLING*

start from the set of un-weighted particles of the previous layer

for $i = 1$ to N **do**

1.1 *SENSOR NOISE*

/* *draw a sample* $\tilde{\mathbf{z}}^{\mathrm{sens},i}$ *from* $p(\mathbf{z}_{gt}^{\mathrm{sens}}\mathbf{z}^{\mathrm{sens}})$ */

for $s = 1$ to N_s **do**

draw sample from von-Mises Fisher $f_p(\boldsymbol{\mu}, \kappa) \rightarrow \mathbf{q}_w$

$\tilde{\mathbf{q}}_s^{TS} = \mathbf{q}_s^{TS} \circ \mathbf{q}_w$

end for

set $\tilde{\mathbf{z}}^{\mathrm{sens},i} = (\tilde{\mathbf{q}}_1^{TS} \dots \tilde{\mathbf{q}}_{N_s}^{TS})^T$

1.1 *INVERSE KINEMATICS*

/* *computation of* $\mathbf{x}_k^{(i)} = g^{-1}(\mathbf{x}_{a,k}^i, \tilde{\mathbf{z}}^{\mathrm{sens}})$ */

for $j = 1$ to N_s **do**

compute: $\mathbf{R}_s^{TS} = \text{quat2mat}(\tilde{\mathbf{q}}_j^{TS})$

compute: $\mathcal{F}(\mathbf{x}_a) \rightarrow \mathbf{R}_j^p, \mathbf{R}_j^c$

set: $\mathbf{R}_j = (\mathbf{R}_j^p)^T \mathbf{R}_s^{TS} \mathbf{R}_s^{SB} (\mathbf{R}_j^c)^T$

solve: $\exp(\theta_{j_1}\widehat{\omega}_{j_1}) \exp(\theta_{j_2}\widehat{\omega}_{j_2}) \exp(\theta_{j_3}\widehat{\omega}_{j_3}) = \mathbf{R}_j$

end for

set: $\pi_k^{(i)} = \exp\left(-\beta_k \cdot V\left(\mathbf{x}_k^{(i)}\right)\right)$

end for

set: $\mathcal{L}_k = \{\pi_k^{(i)}, \mathbf{x}_{a,k}^{(i)}\}_{i=1}^N$

2. *RESAMPLING*

draw N samples from $\mathcal{L}_k \rightarrow \boldsymbol{x}_{a,k}^{(i)}$

3. *DIFFUSION*

$\mathbf{x}_{a,k+1}^{(i)} = \mathbf{x}_{a,k}^{(i)} + \boldsymbol{B}_k$ $\quad \{\boldsymbol{B}_k$ is a sample from $\mathcal{N}(0, \Sigma_a)\}$

end for

[8] Since the basic Gaussian process does not take the correlation of the output variables into account the process is equivalent to a 3rd order regression from previous full state estimates to the manifold coordinates

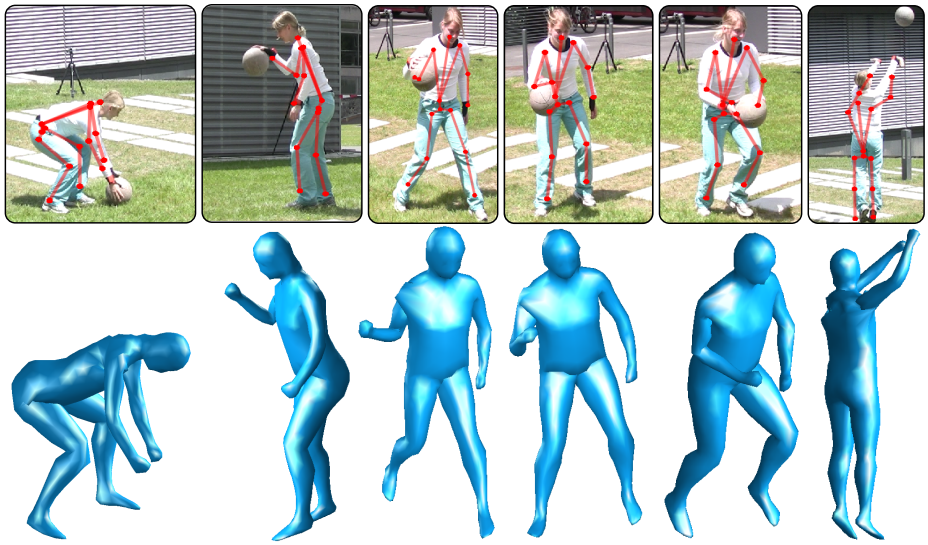


Fig. 10: Tracking with strong illumination

5 Experiments

The standard benchmark for human motion capture is *HumanEva* that consists of indoor sequences. However, no outdoor benchmark data comprising video as well as inertial data exists for free use yet. Therefore, we recorded eight sequences of two subjects performing four different activities, namely walking, karate, basketball and soccer. Multi-view image sequences are recorded using four unsynchronized off-the-shelf video cameras. To record orientation data, we used an Xsens Xbus Kit [36] with 10 sensors. Five of the sensors, placed at the lower limbs and the back, were used for tracking, and five of the sensors, placed at the upper limbs and at the chest, were used for validation. As for any comparison measurements taken from sensors or marker-based systems, the accuracy of the validation data is not perfect but is useful to evaluate the performance of a given approach. The eight sequences in the data set comprise over 3 minutes of footage sampled at 25 Hz. Note that the sequences are significantly more difficult than the sequences of *HumanEva* since they include fast motions, illumination changes, shadows, reflections and background clutter. For the validation of the proposed method, we additionally implemented five baseline trackers: two video-based trackers based on local (L) and global optimization (G) respectively and three hybrid trackers that also integrate orientation data: local optimization (LS), global optimization (GS) and rejection sampling (RS) which we briefly describe here

- (L): Local optimization tracker. The model is projected to the image to find correspondences between the image silhouette contours and the model points. Then, the non-linear least squares problem is solved using a variant of *Levenberg-Marquardt algorithm*, see [15, 25] for more details.

- (G): Global Particle based optimization. Optimization here is performed by means of simulated annealing, *i.e.* , pose hypotheses are generated and weighted with progressively smooth versions of the image likelihood. The final pose is obtained as the average of the particle set in the last annealing layer, see [6, 10] for more details.
- (LS): Local optimization + inertial Sensors. Optimization is again performed by means of non-linear least squares but the cost function to be minimized consists of an image term and a term that models the likelihood of the inertial sensor measurements
$$V(\mathbf{x}) = \mu_1 V^{\mathrm{im}}(\mathbf{x}) + \mu_2 V_1^{\mathrm{sens}}(\mathbf{x})$$
where $V_1^{\mathrm{sens}}(\mathbf{x})$ is defined as the squared Frobenious norm between the sensor and the tracking bone orientation matrices. Both the model-image Jacobian and the orientational Jacobian are derived analytically for better accuracy and efficiency. The algorithm is the based on [24].
- (GS): Global particle based optimization with Sensors. Like the (G) method but including the inertial sensor measurements in the weighting function. We optimize a cost function
$$V(\mathbf{x}) = \mu_1 V^{\mathrm{im}}(\mathbf{x}) + \mu_2 V_2^{\mathrm{sens}}(\mathbf{x})$$
where the image term $V^{\mathrm{im}}(\mathbf{x})$ is the one defined in Eq. (36) and is chosen to be to be a piece-wise increasing linear function of the angular error between the tracking and the sensor bone orientations. That is, for angular errors bigger than 10 degrees we scale the cost by a factor of 5. Big deviations from the orientation measurement could in principle be penalized with a quadratic function but this yields to many particles being rejected in early stages and results in lower performance. Note that although $\mu_2 V_2^{\mathrm{sens}}(\mathbf{x})$ and $\mu_2 V_1^{\mathrm{sens}}(\mathbf{x})$ are not identical they are both functions of distance metrics for rotations and are thus equivalent. For (LS) we optimize $\mu_2 V_1^{\mathrm{sens}}(\mathbf{x})$ because derivatives are easier to compute. We hand tuned the influence weights μ_1, μ_2 to obtain the best possible performance.
- (RS): Rejection Sampling. This method is commonly used to sample hypotheses that satisfy a set of constraints. The method works by sampling hypotheses and rejecting hypotheses that do not satisfy the constraints up to a certain tolerance. It was for example used in [17] to integrate object interaction constraints. For our problem, to combine inertial data with video images we draw particles directly from $p(\mathbf{x}_t|\mathbf{z}^{\mathrm{sens}})$ using a rejection sampling scheme. In our implementation of (RS), we reject a particle when the angular error for any of the constraints is bigger than 10 degrees.

For a comprehensive overview of model based methods for human pose estimation we refer the interested reader to [26].

Let the *validation set* be the set of quaternions representing the sensor bone orientations *not* used for tracking as $\mathbf{v}^{\mathrm{sens}} = \{\mathbf{q}_1^{val}, \ldots, \mathbf{q}_5^{val}\}$. Let $i_s, s \in \{1 \ldots t\}$ be the corresponding bone index, and $\mathbf{q}_{i_s}^{TB}$ the quaternions of the tracking bone orientation (Sect. 3.2). We define the *error measure* as the average geodesic angle between the sensor bone orientation and the tracking orientation for a sequence of T frames as

$$d_{quat} = \frac{1}{5\,T} \sum_{s=1}^{5} \sum_{t=1}^{T} \frac{180^\circ}{\pi} 2 \arccos \left| \left\langle \mathbf{q}_s^{val}(t), \mathbf{q}_{i_s}^{TB}(t) \right\rangle \right| . \tag{40}$$

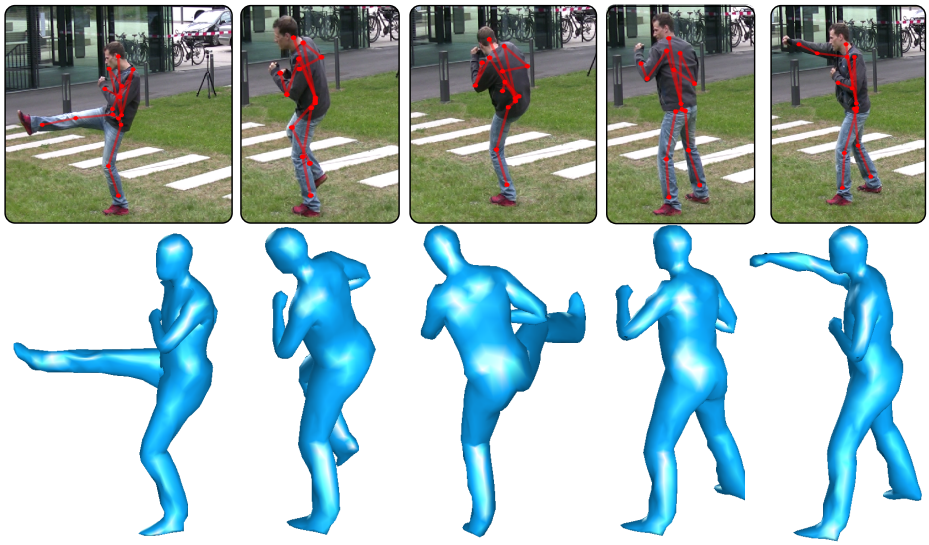


Fig. 11: Tracking results of a karate sequence

Comparison with video and local trackers: We compare the performance of four different tracking algorithms using the distance measure, namely (L), (G), (LS) and our proposed approach (P). We show d_{quat} for the eight sequences and each of the four trackers in Fig. 12. For (G) and (P) we used the same number of particles $N = 200$. As it is apparent from the results, local optimization is not suitable for outdoor scenes as it gets trapped in local minima almost immediately. Our experiments show that LS as proposed in [24] works well until there is a tracking failure in which case the tracker recovers only by chance. Even using (G), the results are unstable since the video-based cues are too ambiguous and the motions too fast to obtain reliable pose estimates. By contrast, our proposed tracker achieves an average error of $10.78\,^\circ \pm 8.5^\circ$ and clearly outperforms the pure video-based trackers and (LS).

Comparison with GS: In Fig. 13 (a), we show d_{quat} for a varying number of particles using the (GS) and our proposed algorithm (P) for a walking sequence.

The error values show that optimizing a combined cost function leads to bigger errors for the same number of particles when compared to our method. This was an expected result since we reduce the dimension of the search space by sampling from the manifold and consequently less particles are needed for equal accuracy. Most importantly, the visual quality of the 3D animation deteriorates more rapidly with (GS) as the number of particles are reduced[9]. This is partly due to the fact that the constraints are not always satisfied when additional error terms guide the optimization.

[9] see the video for a comparison of the estimated motions at http://www.tnt.uni-hannover.de/∼pons/

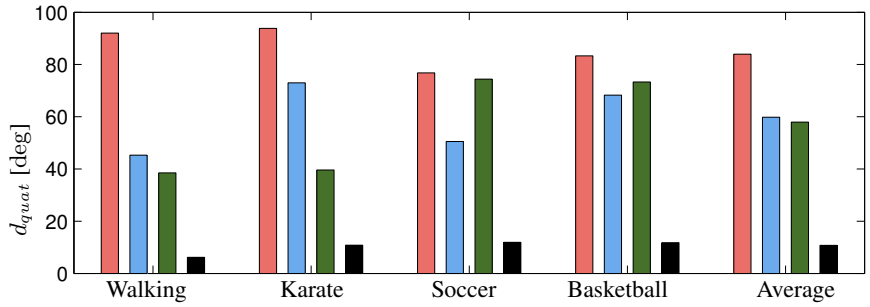


Fig. 12: Mean orientation error of our 8 sequences (2 subjects) for methods (bars left to right) L (local optimization), LS (local+sensors), GL (global optimization), and ours P.

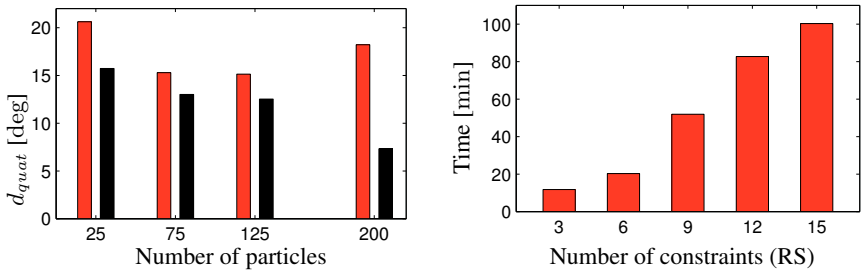


Fig. 13: **(a)**: Orientation error with respect to number of particles with (red) the GS method and (black) our algorithm. **(b)**: Running time of *rejection sampling* (RS) with respect to number of constraints. By contrast our proposed method takes 0.016 seconds for 15 DoF constraints. The time to evaluate the image likelihood is excluded as it is independent of the algorithm.

Comparison with Rejection Sampling (RS): Another option for combining inertial data with video images is to draw particles directly from $p(\mathbf{x}_t|\mathbf{z}^{\text{sens}})$ using a simple rejection sampling scheme. In our implementation of (RS), we reject a particle when the angular error is bigger than 10 degrees. Unfortunately, this approach can be very inefficient especially if the manifold of poses that fulfill the constraints lies in a narrow region of the parameter space. This is illustrated in Fig. 13 (b) where we show the processing time per frame (excluding image likelihood evaluation) using 200 particles as a function of the number of constraints. Unsurprisingly, rejection sampling does not scale well with the number of constraints taking as much as 100 minutes for 15 DoF constraints imposed by the 5 sensors. By contrast, our proposed sampling method takes in the worst case (using 5 sensors) 0.016 seconds per frame. These findings show that sampling directly from the manifold of valid poses is a much more efficient alternative.

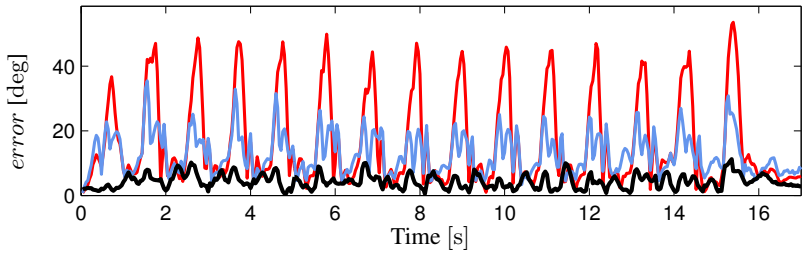


Fig. 14: Angular error for the left hip of a walking motion with (red) no sensor noise model (NN), (blue) Gaussian noise model (GN) and (black) our proposed (MFN).

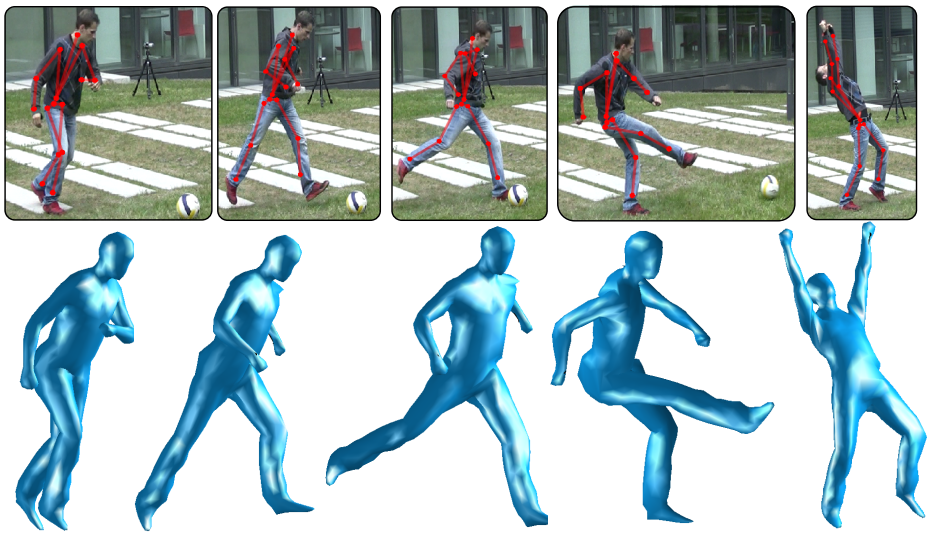


Fig. 15: Tracking results of a soccer sequence

Sensor Noise Model: To evaluate the influence of the sensor noise model, we tracked one of the walking sequences in our dataset using no noise (NN), additive Gaussian noise (GN) in the passive parameters and noise from the von Mises-Fisher (MFN) distribution as proposed in Sect. 4.3. In Fig. 14 we show the angular error of the left hip using each of the three methods. With (NN) error peaks occur when the left leg is matched with the right leg during walking, see Fig. 7. This typical example shows that slight misalignment (as little as $5^\circ - 10^\circ$) between video and sensor data can miss-guide the tracker if no noise model is used. The error measure was 26.8° with no noise model, 13° using Gaussian noise and 7.3° with the proposed model. The error is reduced by 43% with (MFN) compared to (GN) which indicates that the von Mises-Fisher is a more suited distribution to explore orientation spaces than the commonly used Gaussian. This last result might be of relevance not only to model sensor noise but to any particle-based HMC approach. Finally, pose estimation results for typical sequences of our dataset are

shown in Fig. 9, 10, 11 and 15. A video of the proposed approach along with tracking results can be found in the authors website [10].

6 Discussion and Limitations

State-of-the-art video trackers, either based on local or global optimization, suffer from 3D ambiguities inherent in video and usually fail to recover from errors. Our experiments reveal that video based pose estimation algorithms benefit from using a set of small IMUs, specially in outdoor scenarios where the image observation models are weak and ambiguous. Nonetheless, combining inertial and video measurements poses a difficult optimization problem that has to be dealt efficiently. Local optimization is fast and accurate in indoor scenarios. However, our findings indicate that to integrate orientation, (LS) is not suited in outdoor scenarios because it suffers from tracking failures that occur frequently. Optimizing a global cost function (GS) is also not the best choice since it yields an optimization in a high dimensional space which is computationally more expensive. In particular, a high number of hypotheses have to be generated since the search space volume is huge. Rejection sampling (RS) is not suited because it scales very poorly with the number of constraints and the computational time grows exponentially. Finally, we showed that the commonly used Gaussian Noise is outperformed by the proposed von Mises-Fisher noise model when it comes to modeling orientation ambiguities. The reason is that spherical sampling in the joint angle domain does not yield spatially spherical joint configurations as opposed to sampling using (MF). Our proposed method overcomes much of the described limitations: on the one hand the search space is explored only in the region that satisfies the constraints, and on the other hand sampling using Inverse Kinematics has a reinitialization power that overcomes tracking failures in many occasions. Unfortunately, the proposed method is limited by the availability of IMUs. Even though the IMUs are very small and we use only five, they are unavailable in several applications such as surveillance or MoCap and scene understanding from video archives. Another issue that requires improvement is robustness to unsynchronization produced by the IMUs lag during fast motions. The performance of our proposed tracker is still affected from such unsynchronization between IMUs and the video cameras. Since IMUs do not provide any positional measurement, our tracker fails when the body limbs (specially the arms) are not detectable due to long term occlusions. Finally, even though we achieve considerable computational gains w.r.t optimizing the full state space, evaluating the image cost function for every sample is still a bottle neck. To further reduce computational time, an option would be to use very few particles *e.g.* 25 and then locally optimize to obtain better accuracy. Although in this work we have presented an algorithm to combine IMUs with video, the ideas shown here are of significant relevance for the computer vision community. Firstly, the Inverse Kinematics sampling scheme can be used to generate pose hypotheses that satisfy a set of kinematic constraints (we leave extensions to positional constraints as interesting future work). Secondly, the proposed sensor noise model can be used in any problem that involves modeling or optimization of rotation elements.

[10] http://www.tnt.uni-hannover.de/∼pons/

7 Conclusions

By combining video with IMU input, we introduced a novel particle-based hybrid tracker that enables robust 3D pose estimation of arbitrary human motions in outdoor scenarios. As the two main contributions, we first presented an analytic procedure based on Inverse Kinematics for efficiently sampling from the manifold of poses that fulfill orientation constraints. Notably, we show how the IK can be solved in closed form by solving smaller Paden-Kahan subproblems. Secondly, robustness to uncertainties in the orientation data was achieved by introducing a sensor noise model based on the von Mises-Fisher distribution instead of the commonly used Gaussian distribution. Our experiments on diverse complex outdoor video sequences reveal major improvements in the stability and time performance compared to other state-of-the art trackers. Although in this work we focused on the integration of constraints derived from IMU, the proposed sampling scheme can be used to integrate general kinematic constraints. In future work, we plan to extend our algorithm to integrate additional constraints derived directly from the the video data such as body part detections, scene geometry or object interaction.

References

1. Azad, P., Asfour, T., Dillmann, R.: Robust real-time stereo-based markerless human motion capture. In: Proc. 8th IEEE-RAS Int. Conf. Humanoid Robots (2008)
2. Baak, A., Rosenhahn, B., Müller, M., Seidel, H.P.: Stabilizing motion tracking using retrieved motion priors. In: ICCV (2009)
3. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: CVPR (2007)
4. Bregler, C., Malik, J., Pullen, K.: Twist based acquisition and tracking of animal and human kinematics. IJCV 56(3), 179–194 (2004)
5. Chen, J., Kim, M., Wang, Y., Ji, Q.: Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. In: CVPR. pp. 2655–2662. IEEE (2009)
6. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: CVPR. vol. 2, pp. 126–133 (2000)
7. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. IJCV 61(2), 185–205 (2005)
8. Fisher, R.: Dispersion on a sphere. Proceedings of the Royal Society of London. Mathematical and Physical Sciences (1953)
9. Fontmarty, M., Lerasle, F., Danes, P.: Data fusion within a modified annealed particle filter dedicated to human motion capture. In: IRS (2007)
10. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. IJCV 87, 75–92 (2010)
11. Gall, J., Yao, A., Van Gool, L.: 2D action recognition serves 3D human pose estimation. In: ECCV. pp. 425–438 (2010)
12. Ganapathi, V., Plagemann, C., Thrun, S., Koller, D.: Real time motion capture using a time-of-flight camera. In: CVPR (2010)
13. Gavrila, D., Davis, L.: 3D model based tracking of humans in action: a multiview approach. In: CVPR (1996)

14. Hartley, R., Zisserman, A.: Multiple view geometry, vol. 642. Cambridge university press Cambridge, UK (2003)
15. Hasler, N., Rosenhahn, B., Thormählen, T., Wand, M., Gall, J., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. In: CVPR. pp. 224–231 (2009)
16. Hauberg, S., Lapuyade, J., Engell-Norregard, M., Erleben, K., Steenstrup Pedersen, K.: Three dimensional monocular human motion analysis in end-effector space. In: EMMCVPR (2009)
17. Kjellstromm, H., Kragic, D., Black, M.J.: Tracking people interacting with objects. In: CVPR. pp. 747–754 (2010)
18. Lee, C., Elgammal, A.: Coupled visual and kinematic manifold models for tracking. IJCV (2010)
19. Lee, M.W., Cohen, I.: Proposal maps driven mcmc for estimating human body pose in static images. In: CVPR. vol. 2 (2004)
20. Lehment, N., Arsic, D., Kaiser, M., Rigoll, G.: Automated pose estimation in 3D point clouds applying annealing particle filters and inverse kinematics on a gpu. In: CVPR Workshop (2010)
21. Moeslund, T., Hilton, A., Krueger, V., Sigal, L. (eds.): Visual Analysis of Humans: Looking at People. Springer (2011)
22. Murray, R., Li, Z., Sastry, S.: A Mathematical Introduction to Robotic Manipulation. CRC Press, Baton Rouge (1994)
23. Paden, B.: Kinematics and control of robot manipulators. Ph.D. thesis (1985)
24. Pons-Moll, G., Baak, A., Helten, T., Müller, M., Seidel, H.P., Rosenhahn, B.: Multisensor-fusion for 3D full-body human motion capture. In: CVPR. pp. 663–670 (2010)
25. Pons-Moll, G., Rosenhahn, B.: Ball joints for marker-less human motion capture. In: WACV. pp. 1–8 (2009)
26. Pons-Moll, G., Rosenhahn, B.: Model-based pose estimation. Visual Analysis of Humans pp. 139–170 (2011)
27. Pons-Moll, G., Baak, A., Gall, J., Leal-Taixe, L., Mueller, M., Seidel, H.P., Rosenhahn, B.: Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In: IEEE International Conference on Computer Vision (ICCV) (nov 2011)
28. Pons-Moll, G., Leal-Taixé, L., Truong, T., Rosenhahn, B.: Efficient and robust shape matching for model based human motion capture. In: DAGM (2011)
29. Salzmann, M., Urtasun, R.: Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In: CVPR (Jun 2010)
30. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: ICCV. pp. 750–757 (2003)
31. Shoemake, K.: Animating rotation with quaternion curves. ACM SIGGRAPH 19(3), 245–254 (1985)
32. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3D human figures using 2D image motion. In: ECCV (2000)
33. Sigal, L., Balan, L., Black, M.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NIPS. pp. 1337–1344 (2008)
34. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3d human tracking. In: CVPR (2003)
35. Tao, Y., Hu, H., Zhou, H.: Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. IJRR 26(6), 607 (2007)
36. Technologies, X.M.: http://www.xsens.com/
37. Urtasun, R., Fleet, D.J., Fua, P.: 3D people tracking with gaussian process dynamical models. In: CVPR (2006)
38. Wang, P., Rehg, J.M.: A modular approach to the analysis and evaluation of particle filters for figure tracking. In: CVPR (2006)

39. Wood, A.: Simulation of the von mises-fisher distribution. Communications in Statistics - Simulation and Computation (1994)
40. Zhang, F., Hancock, E.R., Goodlett, C., Gerig, G.: Probabilistic white matter fiber tracking using particle filtering and von mises-fisher sampling. Medical Image Analysis 13(1), 5–18 (2009)